Institut de Recherche
pour le Développement
FRANCE

NGS sequencing

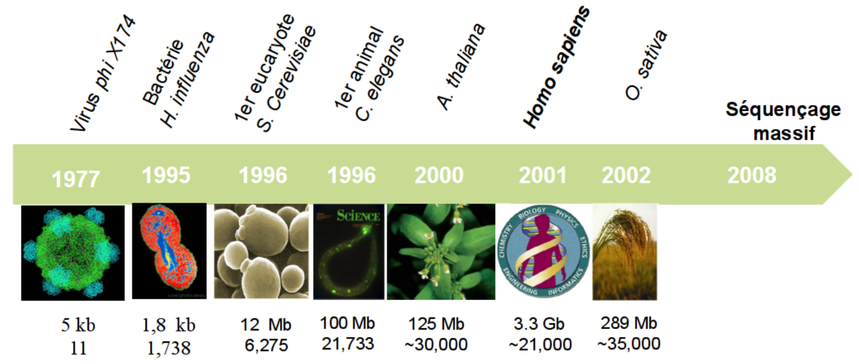
Dr Francois Sabot & Christine Tranchant-Dubreuil

8th of October, 2018

IRD - UMR DIADE

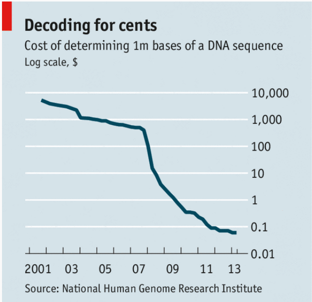
Introduction

A little history of sequencing...



Virus phi X174	Bactérie *H. influenza*	1er eucaryote *S. Cerevisiae*	1er animal *C. elegans*	*A. thaliana*	***Homo sapiens***	*O. sativa*	**Séquençage massif**
1977	1995	1996	1996	2000	2001	2002	2008
5 kb 11	1,8 kb 1,738	12 Mb 6,275	100 Mb 21,733	125 Mb ~30,000	3.3 Gb ~21,000	289 Mb ~35,000	

...From Data Rarity to Data Deluge

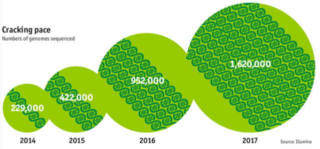


Decoding for cents

Cost of determining 1m bases of a DNA sequence

Log scale, $

Source: National Human Genome Research Institute

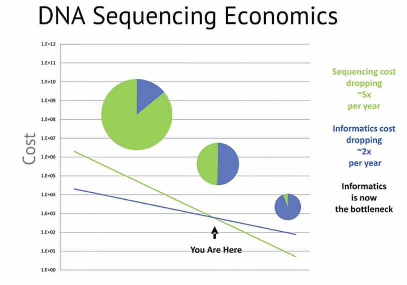


Cracking pace

Numbers of genomes sequenced

Source: Illumina

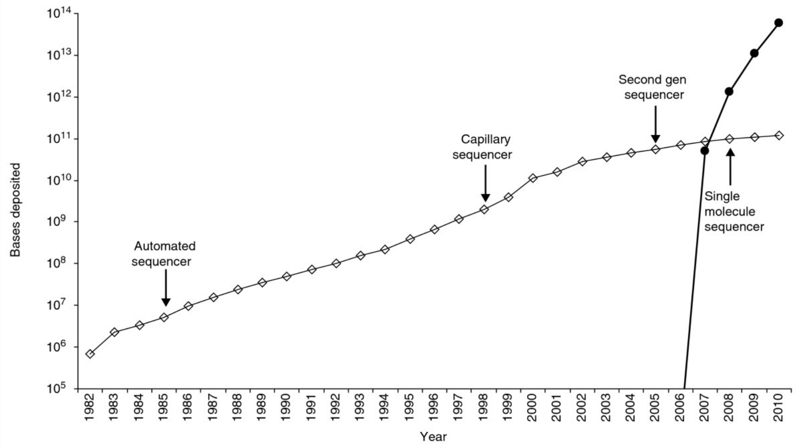
From The economist



DNA Sequencing Economics

From Bussiness Insider

...From Data Rarity to Data Deluge



What can we do with it ?

- Genetic diversity

What can we do with it ?

- Genetic diversity
- Gene discovery

What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure

What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection

What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection
- Metagenomic

What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection
- Metagenomic
- Pangenomic

What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection
- Metagenomic
- Pangenomic
- And many other things...

Methods

Current Technologies

2^{nd} Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

Current Technologies

2^{nd} Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

3^{rd} Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

Current Technologies

2^{nd} Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

454

IonTorrent

Illumina

3^{rd} Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

Current Technologies

2^{nd} Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

454

IonTorrent

Illumina

3^{rd} Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

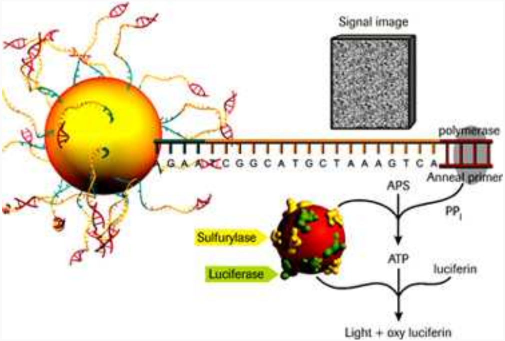
PacificBiosciences

Oxford Nanopore

454

GS Junior
Roche
454 LIFE SCIENCES
Genome Sequencer FLX
Roche

454


Signal image
polymerase
G A A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PP$_i$
Sulfurylase
ATP
luciferin
Luciferase
Light + oxy luciferin

454 - Advantages and Limits

Advantages : Length (400 - 750 bases)

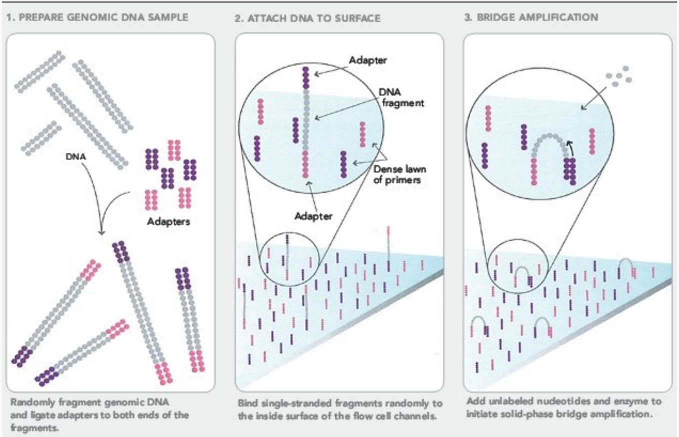
Limits :

- Homopolymers
- Error rate (15%, non random)
- Output volume
- Price

Roche has stopped 454 dev, dying technology

Illumina

Illumina



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

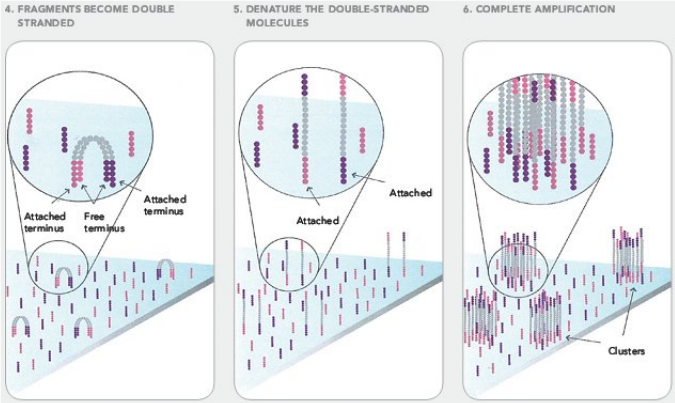
2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Illumina



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

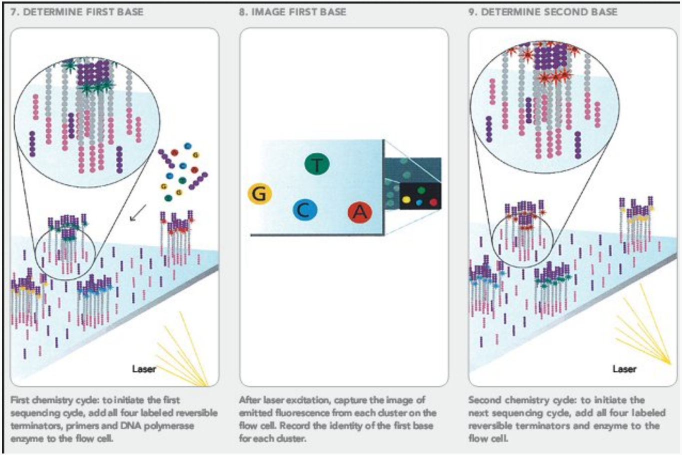
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Illumina



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

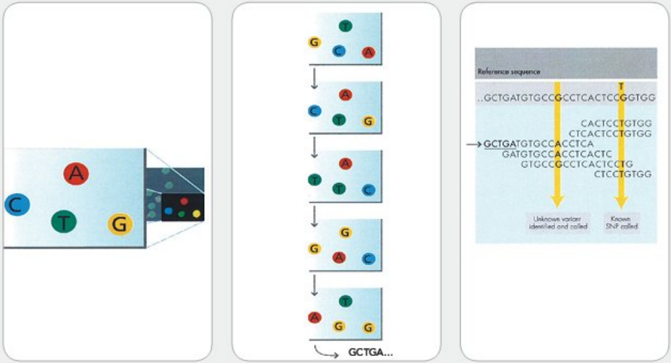
8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Illumina



10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

Illumina - Advantages and Limits

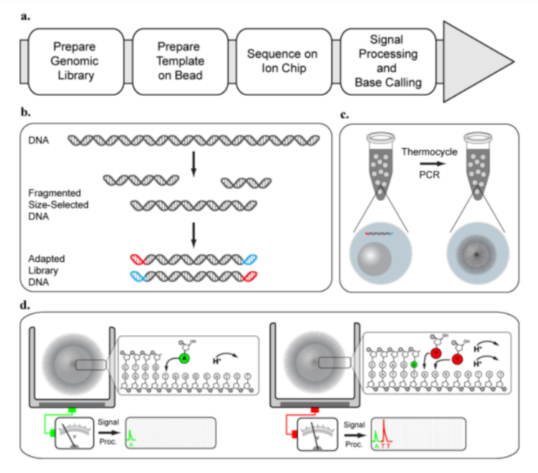
Advantages :

- Output volume (200+ millions of 150b reads, HiSeq 2500)
- Accuracy (99.99 %)
- Run is cheap
- MySeq is cheap (around 60 000 USD per machine)

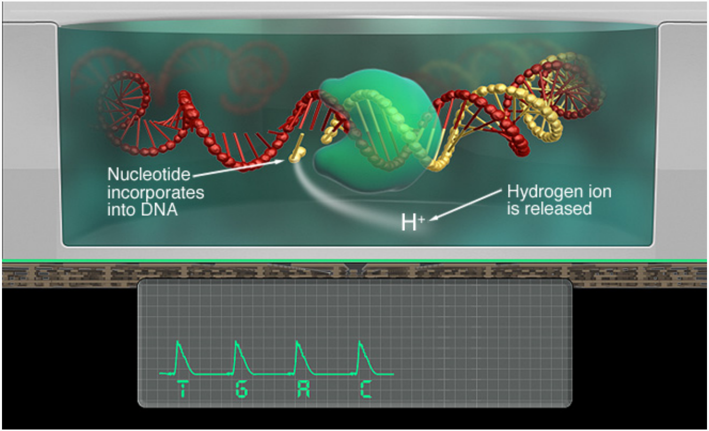
Limits : Size (150 + 150 in HiSeq4000 and X, but 400 for MySeq)

IonTorrent

IonTorrent


a.
Prepare Genomic Library
Prepare Template on Bead
Sequence on Ion Chip
Signal Processing and Base Calling
b.
DNA
Fragmented Size-Selected DNA
Adapted Library DNA
c.
Thermocycle
PCR
d.
H+
Signal
Proc.
A
Signal
Proc.
A T T

IonTorrent


Nucleotide incorporates into DNA
Hydrogen ion is released
H+
T
G
A
C

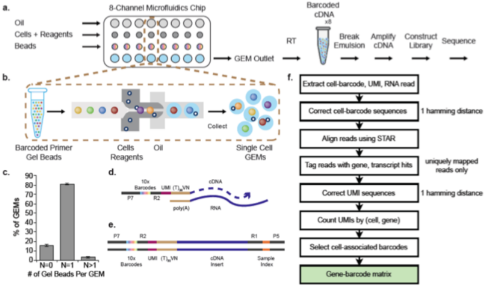
IonTorrent - Advantages and Limits

Advantages :

- Price (less than 200 USD per run)
- Lab sized machine (around 80 000 USD per machine)

Limits : Error rate (15%)

10x Genomics



10x Genomics - Advantages and Limits

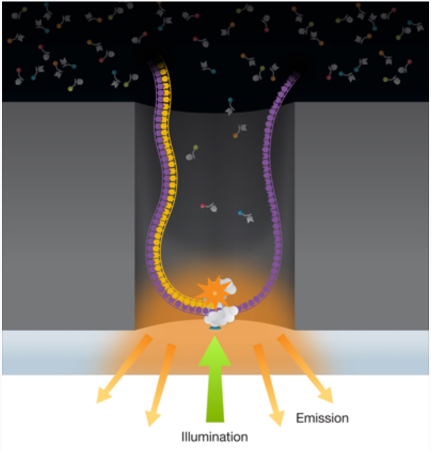
Advantages :

- Price: around 5,000 USD/run (machine at 125k USD)
- Illumina quality data

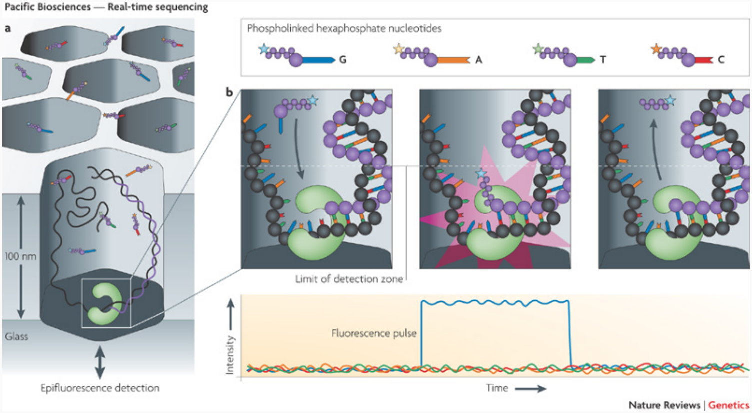
Limits : “Limited” length of long reads (10kb max)

Pacific BioSciences

Pacific BioSciences



Pacific BioSciences


Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Intensity
Fluorescence pulse
Time
Nature Reviews | Genetics

PacBio - Advantages and Limits

Advantages :

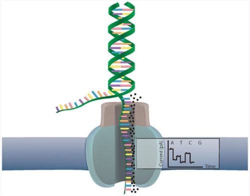
- Length (mean 10kb, more than 40kb regularly)
- Single strand direct sequencing, no amplification bias

Limits :

- Error Rate (15%, but can be corrected)
- Machine size and price (more than 900 000 USD)
- Run price (600 USD for 500 Mb)

Oxford Nanopore

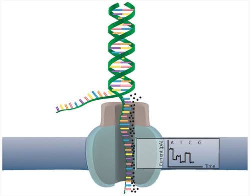
Oxford Nanopore



From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

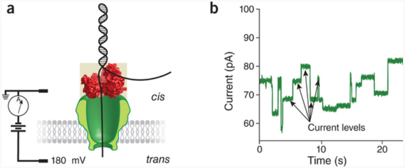
Oxford Nanopore



From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

- Magnetic fields variation measure
- *Minion*: USB key - sized



From Nature Biotechnology

ONT - Advantages and Limits

Advantages :

- Length (mean 10-50kb, more than 2Mb reported)
- Bases Modification detection in real-time
- Single strand direct sequencing
- Machine cheap (2,000 USD for Minion)
- Run cheap (1,000 USD for 30Gb by now minimum)
- Fast: 15mn library, 48-72h run

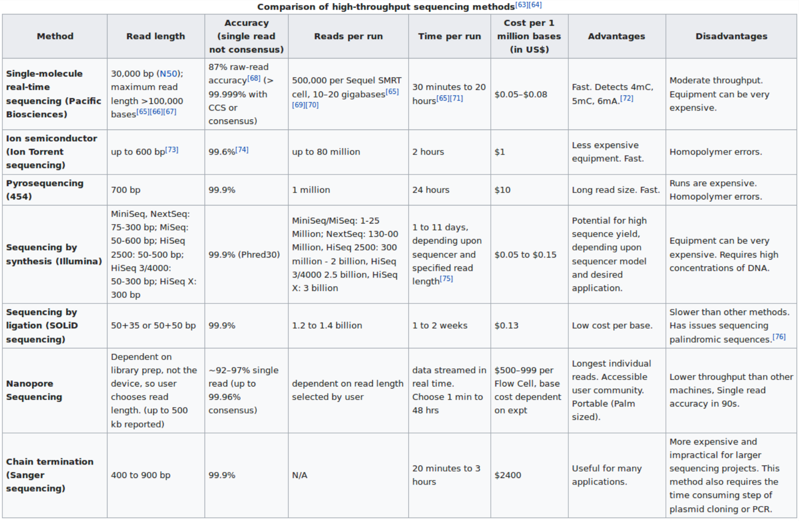
Limits :

- Error Rate (6-9%, but can be corrected)
- Quality of DNA limits the sequencing
- Heu...

Dead-born, old or unusual Technologies

- SOLiD †, because of too small sequence size and no new dev.
- Helicos †, because of too many errors and trouble in chemistry
- Polonator †, because of too small sequence size
- DNA nanoball sequencing (Complete Genomics©), nobody uses it but CG group.
- Single Sequence magnetic bead (ongoing development)
- Transmission electron microscopy DNA sequencing (ongoing development)

Various Technologies, length, cost, outputs...



Comparison of high-throughput sequencing methods[63][64]

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	30,000 bp (N50); maximum read length >100,000 bases[65][66][67]	87% raw-read accuracy[68] (> 99.999% with CCS or consensus)	500,000 per Sequel SMRT cell, 10–20 gigabases[65][69][70]	30 minutes to 20 hours[65][71]	$0.05–$0.08	Fast. Detects 4mC, 5mC, 6mA.[72]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 600 bp[73]	99.6%[74]	up to 80 million	2 hours	$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[75]	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences.[76]
Nanopore Sequencing	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)	~92–97% single read (up to 99.96% consensus)	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$500–999 per Flow Cell, base cost dependent on expt	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2400	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time consuming step of plasmid cloning or PCR.

From Wikipedia

Sample types - Classical

- DNA from plant, animal, microbial...

Sample types - Classical

- DNA from plant, animal, microbial...
- RNA from various sources

Sample types - Classical

- DNA from plant, animal, microbial...
- RNA from various sources
- smallRNA, idem

Sample types - Classical

- DNA from plant, animal, microbial...
- RNA from various sources
- smallRNA, idem
- Environmental sample

Alternative Samples

- Organite DNA (mitochondria, chloroplast)

Alternative Samples

- Organite DNA (mitochondria, chloroplast)
- Subsample RNA (exon capture, 16S capture for Barcoding)

Alternative Samples

- Organite DNA (mitochondria, chloroplast)
- Subsample RNA (exon capture, 16S capture for Barcoding)
- Viral sample from infected tissue

Alternative Samples

- Organite DNA (mitochondria, chloroplast)
- Subsample RNA (exon capture, 16S capture for Barcoding)
- Viral sample from infected tissue
- As many as you can extract...

Thanks for your attention